DOCUMENT RESUME

ED 275 199 FL 016 146

AUTHOR Stansfield, Charles W.
TITLE A History of the Test of Written English: The Developmental Year.
INSTITUTION Educational Testing Service, Princeton, N.J.
PUB DATE 86
NOTE 20p.; Paper presented at an International Invitational Conference on Research in Language Testing (Kiryat Anavim, Israel, May 10-13, 1986).
PUB TYPE Historical Materials (060) -- Speeches/Conference Papers (150) -- Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Educational History; *English (Second Language); Essay Tests; Foreign Students; Higher Education; *Language Tests; *Standardized Tests; *Test Construction; Test Format; Test Items; Test Use; *Writing Evaluation; *Writing Skills
IDENTIFIERS Test of English as a Foreign Language; *Test of Written English

ABSTRACT

A history of the Test of Written English (TWE), a section of the Test of English as a Foreign Language (TOEFL), describes its inception and development process. The new test is a thirty-minute essay test providing a measure of a non-native English-speaker's ability to perform academic writing tasks similar to those required of international students in North American universities. The article describes the impetus for and early concerns about a standardized writing test, preliminary surveys of administrators and teachers of English as a second language, an investigation of the extent to which the existing test assessed academic writing skills, research on academic writing skill needs, a subsequent survey concerning support in the profession for a new writing test, and the development of the instrument itself. The instrument's development includes a continuing process for drafting, selecting, and polishing essay topics, a pretest, determination of the procedures for reading and scoring essays (including reader selection and training, and scoring reliability), solving technological problems associated with changing the overall test format, incorporating TWE scores into the TOEFL scale, and giving the test a name. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED275199

A History of the Test of Written English: The Developmental Year

by

Charles W. Stansfield

Director, Test of Written English

Educational Testing Service

Princeton, New Jersey

This paper was read at LT + 25, an international invitational conference on research in language testing held in honor of the retirement of John Carroll and Robert Lado at Kiryat Anavim, Israel, on May 10-13, 1986. The paper has been accepted for publication in Language Testing.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improver
EDUCATIONAL RESOURCES INFORMAT
CENTER (ERIC)

X This document has been reproduce
received from the person or organiz
originating it.

☐ Minor changes have been made to imp
reproduction quality.

• Points of view or opinions stated in this c
ment do not necessarily represent of
OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Charles Stansfield

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

FL016146

BEST COPY AVAILABLE

Copyright © 1986 by Educational Testing Service. All rights reserved.

Unauthorized reproduction in whole or in part is prohibited.

TOEFL is a trademark of Educational Testing Service, registered in the U.S.A. and in many other countries.

A History of the Test of Written English: The Developmental Year

The Test of Written English (TWE) is a new section of the Test of English as a Foreign Language (TOEFL) that will be offered three times a year beginning in July 1986. This thirty-minute essay test provides a measure of a nonnative English speaker's ability to perform academic writing tasks similar to those required of international students in North American universities. (For a description of the TWE program, see Stansfield and Webster, 1986.)

In order to describe adequately the history of the Test of Written English, one must describe its prehistory. Since its first administration in 1964, the TOEFL examination has been an indirect measure of English language proficiency. The test included five multiple-choice sections until September 1976, at which time the five sections were combined into three (Angelis, 1979). These three sections provide separate measures of listening comprehension, knowledge of structure and written expression, and reading comprehension and vocabulary. Although previous research has shown that performance on the structure and written expression items (section 2) correlates highly with scores on direct measures of writing ability (Pitcher and Ra, 1967; Pike, 1979), many TOEFL score users have questioned the validity of this section as a measure of a nonnative speaker's ability to write in a college classroom. The results of several studies conducted during the last five years also reflect these doubts.

At the request of the TOEFL Research Committee,[1] Hale and Hinofotis (1981) interviewed twenty-five leaders of the teachers of English as a second langauge (ESL) profession in North America to identify new trends in language assessment practice. When asked how TOEFL could do a better

job of preadmission assessment, many recommended the inclusion of a direct writing test. The researchers noted: "Some felt that an objective test could serve a useful screening function; others indicated great reluctance to make decisions about writing ability from an indirect measure, expressing concern about lack of face validity" (p. 7).

A TOEFL research study by Angelis (1982) reached the same general conclusion. Angelis surveyed graduate faculty members in engineering and business, the two fields that enroll the largest numbers of international students. Both groups indicated that the writing deficiencies of these students were a major concern, and many respondents indicated that TOEFL does a limited job of providing information about productive skills such as writing. They noted that there may be little relationship between knowledge of correct written expression, as indicated by multiple-choice tests, and actual writing skills.

At the request of the TOEFL program office, Kane (1983) surveyed administrators and ESL teachers at 600 institutions in the United States. In response to the question, "If you could redesign TOEFL to provide optimum utility for your program, what changes would you make?" he found that the most frequently suggested change was the inclusion of a direct measure of writing.

Carlson, Kline, and Ward (1984) surveyed 194 community colleges with substantial international student enrollments in an effort to gain a better understanding of their testing practices. Respondents to a questionnaire assigned the highest priority to the assessment of international students' reading and writing skills for placement purposes. For writing assessment, most respondents preferred a direct format.

Because the investigations by Hale and Hinofotis (1981) and Angelis (1982) documented concerns about whether TOEFL provided information about productive writing skills, it became necessary to determine the extent to which the three-section version of TOEFL is a valid indicator of the academic writing skills required of undergraduate and graduate applicants to institutions. However, before a meaningful validation study could be conducted, academic writing skills needed to be defined. To that end, Bridgeman and Carlson (1983) surveyed faculty in 190 departments at 34 large universities in the United States and Canada. They found that short lab reports and brief articles and summaries are common writing assignments in engineering and the natural sciences, while longer research papers are commonly assigned to beginning undergraduates in general, to graduate students in business, and to students in some engineering and science programs.

Bridgeman and Carlson also found that, because of the different writing tasks assigned, professors in different fields do not agree about the types of writing skills that prospective students should be asked to demonstrate. A test of descriptive writing was seen as sufficient in computer science and in some engineering fields. In contrast, the ability to argue for a particular position was seen as important for undergraduates and MBA students, but of limited importance in some engineering and science fields. These departments preferred a writing sample that require examinees to describe or interpret a graph or a chart.

With this information about academic writing tasks, researchers at Educational Testing Service (ETS) proceeded to validate the current TOEFL

against productive writing skills. Carlson, Bridgeman, Camp, and Waanders (1985) obtained four thirty-minute writing samples from each of 638 applicants for admission to U.S. universities as undergraduate and graduate students in business, engineering, and the social sciences. Two of the writing topics required the examinee to compare and contrast different viewpoints and take a position; the other two involved interpretation of a chart or graph. The writing samples were scored by twenty-three English composition and rhetoric instructors and twenty-three ESL instructors, with each essay being scored by at least one reader of each type. Subsamples were scored by graduate professors in engineering and the social sciences. Each sample was rated holistically (general impression as to overall quality), and, during a separate rating session, two of the samples from each student were assigned separate scores for sentence level skills and discourse level skills. TOEFL scores were also obtained for all the applicants.

Correlations and factor analyses of the various scores showed that, while the writing samples and TOEFL were related, each also measured something that was not assessed by the other. In examining other findings and their implications for a potential new measure of writing skills, I find it noteworthy that (a) holistic scores, discourse level scores, and sentence level scores were very closely related, suggesting that a holistic score alone should be an adequate representation of the examinee's writing skills; (b) correlations among topics were as high across topic types as within topic types (In my opinion, this finding suggests that the two topic types assessed the same construct. Thus, for purposes of construct validity, only a single type of writing need be measured.); (c) scores of raters who were ESL instructors, English instructors, and instructors of other disciplines

were all highly intercorrelated, suggesting that individuals in all groups could be trained readily to score using common criteria.

On being informed of the results of these studies in November 1984, the TOEFL Policy Council authorized the development of a direct test of writing (Stansfield, 1984). The first stage of this development included a survey of admissions officers at more than 800 institutions that receive a large number of TOEFL scores each year. The results of the survey, based on a 73 percent return, demonstrated broad support for the proposed TOEFL writing test (Adams-Fallon & Stansfield, 1985). Approximately 80 percent of the respondents favored its establishment. Community colleges, which traditionally have an open-door enrollment policy and do not use admissions tests, were as interested in the new measure as were four-year colleges.

In addition to expressing support, the respondents answered a number of questions about major design features of the proposed test. Approximately 75 percent felt that the writing sample should be compulsory. A similar percentage said the writing sample should be a response to a general rather than a discipline-specific topic, and an even larger number (86 percent) recommended that the writing test score be reported separately, rather than incorporated into the score for structure and written expression. Fifty-six percent preferred a numerical score on the test, as opposed to a letter grade (8 percent) or a descriptive statement of the applicant's writing skills (36 percent).

Early in the project, it was decided that the development of successful essay topics requires the expertise of teaching professionals who have extensive classroom experience and familiarity with the current population

of students. Thus, TOEFL staff enlisted the assistance of seven academics with sound theoretical and practical experience in essay testing. This group, known as the Core Readers, was given extensive responsibility for the development and scoring of the test. They were told that they would be responsible for drafting, selecting, and polishing the essay topics to be pretested, that they would evaluate pretest topics after reading pretest essays obtained by ETS staff, and that they would be responsible for the training of readers and the scoring of essays written at official TOEFL administrations.

The first meeting of the Core Readers was held in Princeton in August 1985. Prior to this meeting, each member of the group was asked to prepare ten possible essay topics. The seventy topics were carefully scrutinized at the meeting, and eleven were deemed to be both appropriate for the TOEFL population and in compliance with racial, cultural, and other sensitivity guidelines used by ETS test development specialists. These topics were selected for pretesting in English language institutes and community colleges in North America and in bilingual schools in other countries. Approximately 200 essays were collected on each topic and read by the Core Readers at their second meeting held in Berkeley, California, in November 1985. Of the eleven topics pretested, six performed very well and were approved for possible operational administration; three were discarded as flawed in some critical way; one was held for revision and pretesting in a revised form; and one was held for minor additional pretesting to ensure the quality of its performance. Two additional meetings of the Core Readers were held early in 1986. These meetings, which usually last three days, will continue to be held four to six times per year.

Another major concern during the developmental year has been the development and validation of a scoring guide. In the study described above, Carlson et al. (1985) trained readers to score papers using a six-point, holistic scale. While points on the scale were defined by sample essays for each topic, no scoring guide was developed. It was decided that for the TOEFL essay test, however, the development and use of a scoring guide would help readers maintain common standards and good reliability. Also, it was felt that care should be taken to anchor this guide in the Carlson et al. essays, since the scores assigned in their study had provided a theoretical base for the TOEFL writing test.

To begin work on the guide, TOEFL staff contracted Kyle Perkins, a professor of linguistics at Southern Illinois University who has published widely on writing assessment, to examine some 200 essays collected by Carlson et al. The particular essays selected were those on which there was greatest agreement among the raters. Each paper had been scored from two to eight times as part of the Carlson et al. study. The papers were grouped by score level, and Perkins was asked to analyze the characteristics of writing in each group. Samples of equal size were drawn from each of the four topics, to ensure that the analysis would be applicable to performance on other topics. After carefully analyzing the lexical, syntactic, and communicative characteristics of the papers in each group, Perkins constructed a 150-250 word description of the strengths and weaknesses that characterized papers in each group. His analyses were submitted to the Core Readers.

A particular problem faced by the TOEFL program is the need for a guide that can be used by essay readers who will rate some thirty-five

essays an hour. The Core Readers felt that such a guide should define each point on the scale in a single statement, and follow this definition with several short, one-line descriptions highlighting specific aspects of discourse that characterize writing at that level.

At the first pretest reading in November 1985, four extra readers, who had extensive essay reading experience and who were not involved in the development of the topics, were invited to assist the Core Readers in rating approximately 2,000 essays that were scheduled to be read as part of the topic evaluation procedure. After spending one and one-half days reading essays on eleven topics, these four readers were asked by the Core Readers to develop a rapid-reference scoring guide based on the analysis provided by Perkins. The guide they produced was then discussed and revised by the Core Readers. Additional minor revisions were made on the guide subsequent to the meeting. The third version of the guide was then submitted to three experts, who used it to rescore the original sample sent to Perkins. A subsequent analysis of their scores showed 80-85 percent agreement with the original scores obtained in the Carlson et al. study. Each expert independently analyzed discrepancies between the two sets of scores to identify any pattern of deviation. However, no pattern was discovered. As a result, all three experts recommended continued use of the third version, although two suggested minor revisions in the wording. The experts who rescored the Carlson essays with the new scoring guide were Kyle Perkins, who had done the original linguistic analysis of the Carlson papers, Barbara Kroll, a professor of ESL at the University of California at Los Angeles who was also one of the four extra readers who developed the guide, and Agnes Yamada, chairperson of the English Depart-

ment at California State University at Dominguez Hills and Chief Reader among the Core Readers.

The Core Readers met again in February 1986 for their second pretest reading. The scoring guide was used at this reading to rate approximately 200 responses to each of eight pretested topics. No revisions appeared necessary to the Core Readers. The group was told that three independent validations of the guide were being carried out and that these would be discussed at a meeting in March. At the March meeting, the Core Readers considered the experts' suggestions for minor revisions. Most of these suggestions were approved, with the result that a fourth draft of the guide was prepared after the meeting. This version was sent by mail to the Core Readers for one final review. No additional changes were suggested.

Through this process, the TOEFL program has developed a guide that can be used rapidly and successfully by dozens of essay readers at operational readings. While it may be necessary to change a word or two during the first year of the program, based on additional experience gained at operational readings of tens of thousands of papers, the integrity of the current guide will remain. This guide will serve to anchor papers on different topics in future years, thereby helping to ensure that a given score will always represent the same degree of writing performance as measured by the test.

During the summer of 1985, TOEFL staff began to consider how the thousands of operational essays would be read. Several alternatives were explored. These included reading the essays at several different locations throughout the United States, reading them at the ETS headquarters

in Princeton, New Jersey, contracting with another company to have the essays read, and reading the essays at the ETS field service office in Berkeley, California. After investigating each alternative, TOEFL staff chose the last.

There were many reasons for this choice. First, the Berkeley office of ETS has a long history of managing essay readings. Over the years, it has managed the reading of the essays produced for the California State University English Placement Test, the NTE Communications Skills test, the South Carolina State Teachers Examination, the PreProfessional Skills Test, the Foreign Service Examination, the California High School Proficiency Examination, and others. As a result, there is a large number of trained readers in the San Francisco Bay Area. The Berkeley office has statistics on the reliability of these readers and thus is in a position to select readers suitable for the TOEFL project. In addition, TOEFL staff was aware that the San Francisco Bay Area has long been a center of activity in writing assessment. The Bay Area Writing Project, which began there twelve years ago, has been emulated in over 100 cities throughout the United States. The National Writing Project is housed at the University of California at Berkeley, as is the National Center for the Study of Writing. Because of this, we concluded that the San Francisco Bay Area, of which Berkeley is a part, contains more teachers who are highly trained in holistic scoring than does any other metropolitan area within the United States. While most of these teachers come from the English composition field, many also teach writing to ESL students.

After the decision was made to hold the essay readings in Berkeley, the issue of reader qualifications arose. TOEFL staff established certain

ground rules to ensure the high quality of readers in the future. It was decided that the most important qualifying criterion should be performance, that is, each reader should have demonstrated that he or she can read reliably and on scale. Therefore, we chose to select only those readers for whom ETS had statistics addressing these issues.

A second consideration in the selection of readers is our desire to have a nearly equal mix of readers from the fields of English composition and ESL. While statistics on more of the former are available, the chairperson of the Core Readers will conduct training sessions in Berkeley for interested ESL teachers who have not read for ETS previously. At these sessions, potential readers will receive five hours of training in holistic scoring using the TWE scoring guide, followed by the uninterrupted reading of thirty papers for which the official scores have already been determined. Again, these papers will be selected from among the essays written for Carlson et al. Each reader's scores will be correlated with the official scores, and only those with the highest correlation will be selected for operational readings. All readers will be experienced teachers of English or ESL at the secondary school or college levels. By including among the readers nearly equal numbers of teachers from both groups. we believe that the scores assigned will be adequately anchored in the quality of writing that one can expect of native speakers, while ensuring that the unique characteristics of the writing produced by ESL learners will be considered.

Progress has been made in a number of other areas this year. Besides having developed enough topics of each type for use during the first operational year, we have made progress in designing specifications for

future topic writers. Although at the start of the program the Core Readers possessed a great deal of experience in the writing of compare-contrast-and-take-a-position topics, this was not the case for chart/graph topics, which, to our knowledge, have not been used previously in large-scale writing assessments. Thus, during the developmental year we have focused our energy on gaining more experience in writing suitable questions of the latter type. We have also tried to preserve this experience by tape-recording the discussions of topics during the pretest readings, and then organizing these insights for the benefit of future Core Readers. Thus, we now have a fifteen-page set of specifications for writers of chart/graph topics. We have not yet produced specifications for the compare-contrast-and-take-a-position topics, but we do not anticipate difficulty in writing these specifications in the future. Indeed, our core readers have found it much easier to write such topics, and those compare/contrast topics that have been pretested this year have been successful a larger percentage of the time than have the chart/graph topics.

We have also solved a number of technological problems relating to the addition of an essay to TOEFL. It was necessary to redesign the TOEFL answer sheet so that, for the administrations that will include an essay, it will contain four sides on two pages and will be perforated in the center. Each page will contain a pregridded number that can be read by an optical scanner. The two portions of the answer sheet, the multiple-choice section and the essay section, will be scanned at different times and at different locations. The multiple-choice portion will be scanned in Princeton, the essay portion in Berkeley. A record that includes the

pregridded number will be constructed for each examinee, and this number will be used to match records from the two files. Data will be transferred from Berkeley to Princeton via telephone lines. An edit routine has been developed to check each examinee's record of reader scores and identification data in Berkeley to determine that the record is complete. The record will be edited again in Princeton after the records from the two files have been merged. Subsequently, scores will be determined and printed on score reports.

One problem we have not yet solved is how to incorporate the essay score into the TOEFL scale. TOEFL makes use of three-parameter item response theory equating (Cowell, 1982, Hicks, 1983), which requires that most questions be pretested on a large sample and placed on an ability scale prior to their inclusion in the test. Thus, it seems that in order for new topics to be equated with old ones, they would have to be administered jointly. TOEFL's essay pretesting is based on a relatively small sample, and each examinee writes on only one topic. It may be possible to equate topics after we have a bank of used ones available for inclusion in the pretesting, but this remains to be seen. The fact that there was no feasible operational solution to the problem of statistically equating essays and placing them on the TOEFL scale was one factor in our decision not to include the essay score in the total TOEFL score at this time. We will continue to work toward a solution, however.

One final detail regarding the name of the test merits mention here. Just one year ago TOEFL staff received approval to begin developing an essay test to be administered worldwide as part of the TOEFL. Because of pressing operational matters, we did not have the lead time necessary to

decide on an appropriate name for the test. Therefore, in the interim we referred to it generically as "the writing test." During the fall of 1985, TOEFL staff reviewed several possible names. Some of these were similar to the names of other ETS tests, so it was necessary to investigate whether any names would generate concern either within ETS or among its clients about their potential to be confused with other existing instruments. In March 1986 we completed the investigation and decided to name the instrument the Test of Written English. This name is linguistically symmetrical and complementary to the TOEFL program's other direct test of language skills, the Test of Spoken English (TSE). Thus, the two names will be easily identifiable as TOEFL tests by users in the field.

During the first year we expect many users to continue to refer to the TWE by its generic name, the writing test. Indeed it is referred to as such in the 1986-87 Bulletin of Information for TOEFL and TSE and on the score report. The name, however, has never been capitalized in any publication, and in the future, the test will be referred to only by its official name, the Test of Written English.

Many decisions have been made during this developmental year that will affect the TWE for years to come. The rationale underlying some of these decisions has been described here. It is hoped that the information contained in this paper will contribute to an improved understanding of the instrument. Comments about the test are welcome and may be addressed to the author.

Charles W. Stansfield
Director, Test of Written English
Educational Testing Service
Princeton, NJ 08541, USA

Notes

[1]A continuing program of research related to TOEFL is carried out under the direction of the TOEFL Research Committee. Its six members include representatives of the TOEFL Policy Council, the TOEFL Committee of Examiners, and distinguished English-as-a-second-language specialists from the academic community. The committee meets twice yearly to approve proposals for test related research and to set guidelines for the entire scope of the TOEFL research program. At the time of the Hale and Hinofotis (1981) study, the members of the TOEFL Research Committee were G. Richard Tucker (Chair), Louis Arena, H. Douglas Brown, Frances Hinofotis, Diane Larsen-Freeman, and David Sparks.

References

Adams-Fallon, M., & Stansfield, C. W. (1985, February). Report to the TOEFL Executive Committee. Internal document. Princeton, NJ: Educational Testing Service.

Angelis, P. J. (1979). TOEFL in recent years. In B. Spolsky (Ed.), Some major tests. Advances in language testing series: 1. Papers in Applied Linguistics. Arlington, VA: Center for Applied Linguistics.

Angelis, P. (1982). Language skills in academic study. Final report submitted to the TOEFL Research Committee. Princeton, NJ: Educational Testing Service. Also published as Academic needs and priorities for testing. American Language Journal, 1, 41-56.

Bridgeman, B., & Carlson, S. (1983). Survey of academic writing tasks required of graduate and undergraduate foreign students. TOEFL Research Report No. 15. Princeton, NJ. Educational Testing Service.

Carlson, S. B., Bridgeman, B., Camp, R., & Waanders, J. (1985). Relationship of admission test scores to writing performance of native and nonnative speakers of English. TOEFL Research Report No. 19. Princeton, NJ: Educational Testing Service.

Cowell, W. R. (1982). Item-response-theory pre-equating in the TOEFL testing program. In P. W. Holland & D. B. Rubin (Eds.), Test equating (pp. 149-161). New York: Academic Press.

Hale, G. A., & Hinofotis, F. (1981). New directions in English language testing. Internal report submitted to the TOEFL Research Committee. Princeton, NJ: Educational Testing Service.

Hicks, M. M. (1983). True score equating by fixed b's scaling: A flexible and stable equating alternative. Applied Psychological Measurement, 7(3), 255-266.

Kane, H. (1983). A study of practices and needs associated with intensive English language programs: Report of findings. Internal report submitted to the TOEFL Program Office. New York: Kane, Parsons and Associates, Inc.

Pike, L. W. (1979). An evaluation of alternative item formats for testing English as a foreign language. TOEFL Research Report No. 2. Princeton, NJ: Educational Testing Service.

Pitcher, B., & Ra, J. B. (1967). The relation between scores on the Test of English as a Foreign Language and ratings of actual theme writing (Statistical Report 67-9). Princeton, NJ: Educational Testing Service.

Stansfield, C. W. (1984, November). Request for funding for writing project. Internal document submitted to the TOEFL Policy Council. Princeton, NJ: Educational Testing Service.

Stansfield, C. W., & Webster, R. (1986, March). The new TOEFL writing test. Paper delivered at the Twentieth Annual TESOL Convention, Anaheim, CA. TESOL Newsletter (in press).